# Turning Big Data into Actionable Information with IBM InfoSphere Streams

## IBM Redbooks Solution Guide

Big data has multiple uses in every industry, from analyzing larger volumes of data than previously possible to driving more precise answers, to analyzing data at rest and data in motion to capture opportunities that were previously lost. As the amount of data that is available to enterprises and organizations increases, more companies are looking to turn this data into actionable information and intelligence in real time. To address these requirements, applications must be able to analyze potentially enormous volumes and varieties of continuous data streams to provide decision makers with critical information almost instantaneously.

By implementing a big data platform, your organization can tackle complex problems that it previously could not solve by using a traditional infrastructure. IBM® InfoSphere® Streams provides a development platform and runtime environment where you can develop applications that ingest, filter, analyze, and correlate potentially massive volumes of continuous data streams, such as those in Figure 1. These data streams are based on defined, proven, and analytical rules that alert you to take appropriate action, all within an appropriate time frame for your organization.

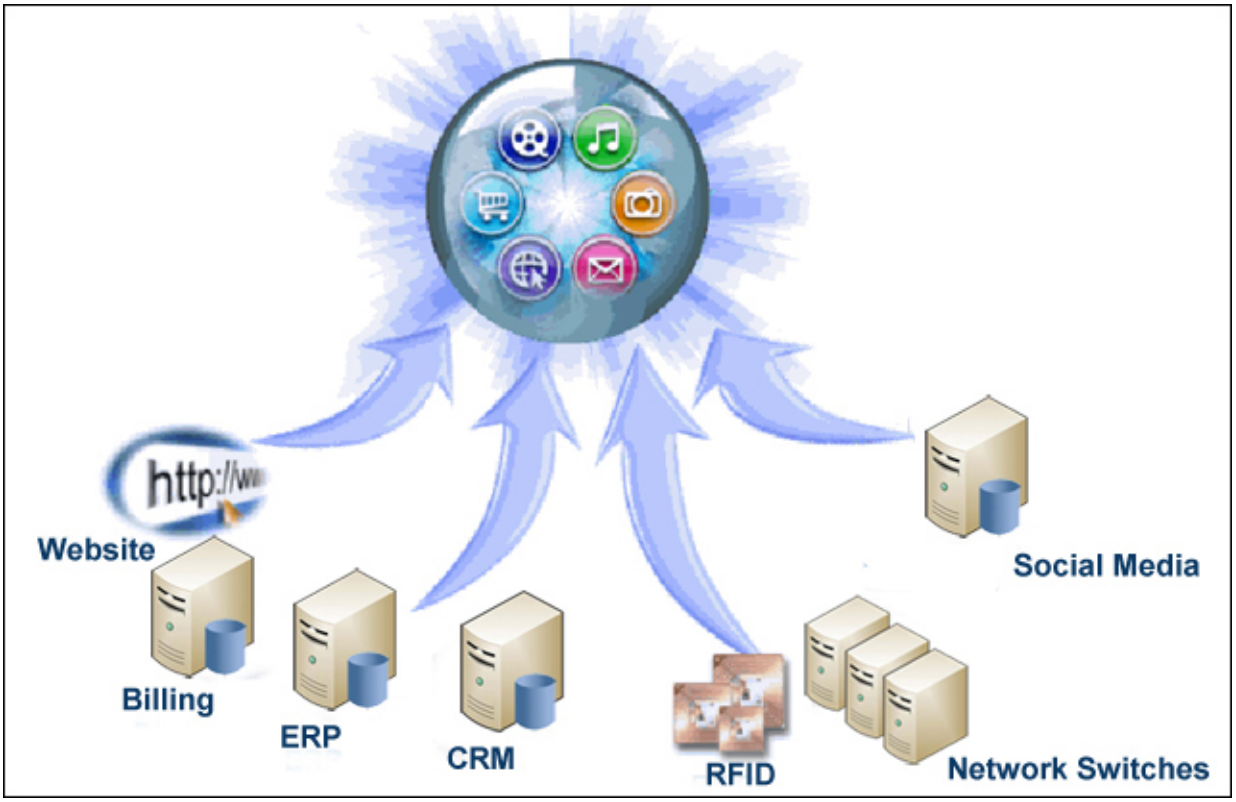


Figure 1. Flow of big data into IBM InfoSphere Streams

## Did you know?

When effectively captured, managed, and analyzed, new information mined from big data can enhance the effectiveness of a country or region's infrastructure. For example, in a world with an expanding population, it can reduce the strain on services and infrastructure, and it can dramatically improve healthcare outcomes with greater efficiency and less investment. In areas of intensified threats to public safety and national borders, this information offers greater levels of security. When more frequent and intense weather events occur, it offers greater accuracy in prediction and management. With more cars, it creates less congestion. As insurance claims rise, it reduces fraud. As natural resources become more scare, it provides more abundant and less expensive energy. The impact of big data has the potential to be as profound as the development of the Internet itself.

## Business value

Typical analytical processes and tools are limited to using stored and (usually) structured data. Data acquisition historically required several time-consuming steps, such as collection through data entry or optical data scanning, cleansing, transformation, enrichment, and finally loading into an appropriate data store. The time it takes to complete these steps results in a delay before the data can be analyzed and used to drive any decisions. Often this delay is enough that any action taken from the analysis becomes more reactive than proactive. You can expect the situation to only get worse. Because our world is becoming more instrumented, traditionally unintelligent devices are now a source of intelligent information. Tiny processors, many with more processing power than the desktop computers of years ago, are being infused in the everyday objects of our lives.

To automate and incorporate streaming data into the decision-making process, you can use a new paradigm in programming called stream computing. Stream computing is the response to the shift in paradigm to harness the potential of data in motion. In traditional computing, you access relatively static information to answer evolving and dynamic analytic questions. With stream computing, you can deploy a static application that continuously applies that analysis to an ever-changing stream of data.

The rise in popularity of social networks indicates a considerable personal value in real-time information. In this forum, millions of people (customers, students, patients, and citizens) are voicing their opinions about everything, from good and bad experiences with products or companies to their support for current issues and trends. Businesses that can analyze this information and that can align with popular desires will be a powerful force.

Businesses are aware that knowing how to effectively use all known sources of data in a way that allows future sources to be incorporated smoothly can be the deciding factor that fuels competitive, economic, and environmental advantages in real time. Those companies and individuals that position themselves to use data in motion (streams) will gain a clear competitive edge. The time it takes to commit the data to persistent storage can be a critical limitation in highly competitive marketplaces. More and more, the effective key decisions need to come from the insights that are available from traditional and nontraditional sources.

## Solution overview

The IBM InfoSphere Streams software platform enables the development and execution of applications that process information in data streams. InfoSphere Streams enables continuous and fast analysis of massive volumes of moving data to help improve the speed of business insight and decision making. InfoSphere Streams can perform analytics on continuously streaming data before it goes to a data warehouse. InfoSphere Streams computing is ideal for high-velocity data where the ability to recognize and react to events in real time is a necessary capability. Although other applications provide stream-computing capabilities, the InfoSphere Streams architecture takes a fundamentally different approach to continuous processing that differentiates it from other platforms. Its distributed runtime platform, programming model, and tools for developing continuous processing applications promote flexibility, development for reuse, and unparalleled performance.

Figure 2 illustrates the flow of the InfoSphere Streams solution.

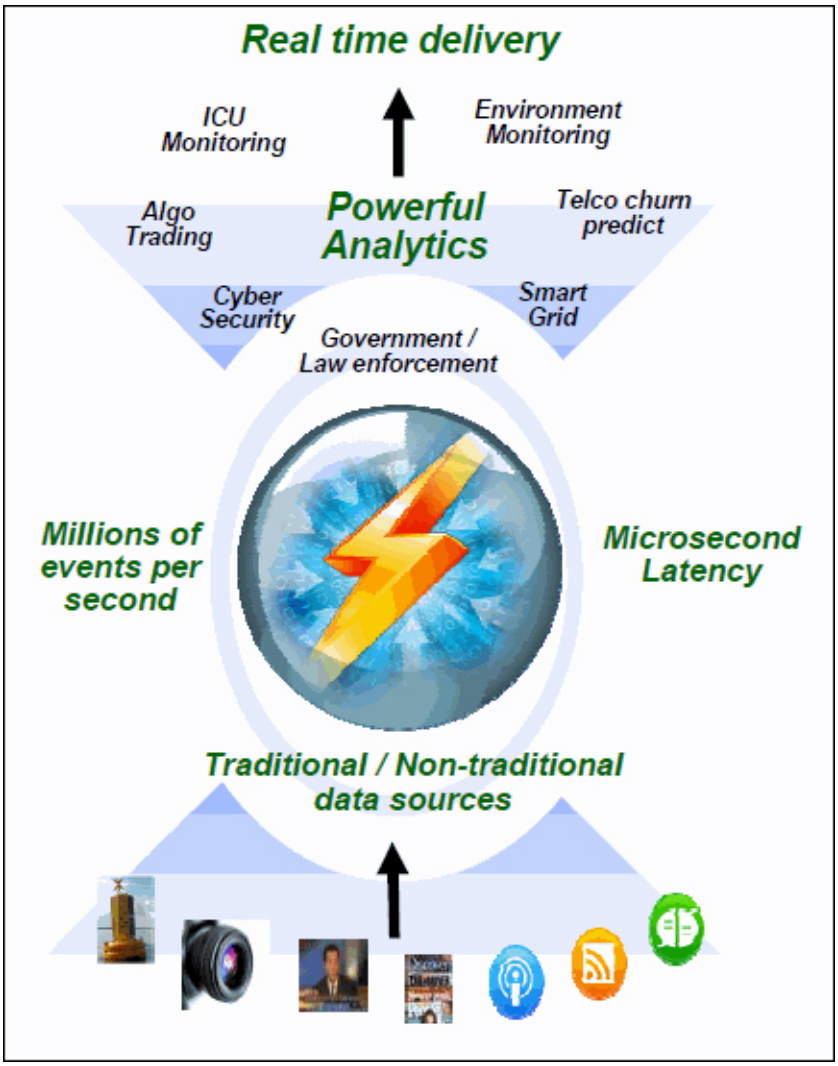

Figure 2. InfoSphere Streams solution

InfoSphere Streams provides simple and complex analytics on continuous data streams. It can scale for computational complexity and supports a wide range of relational and nonrelational data types. InfoSphere Streams supports high data rates and a broad range of data types. For example, data sources that are consumable by InfoSphere Streams include data from sensors, cameras, video, audio, sonar or radar inputs, news feeds, stock tickers, and relational databases.

## Solution architecture

Standard database servers generally have a static data model, data, and dynamic (often long running) queries. IBM InfoSphere Streams supports a widely dynamic data model and data. The InfoSphere Streams version of a query runs continuously without change. Figure 3 illustrates these two approaches to data analysis.

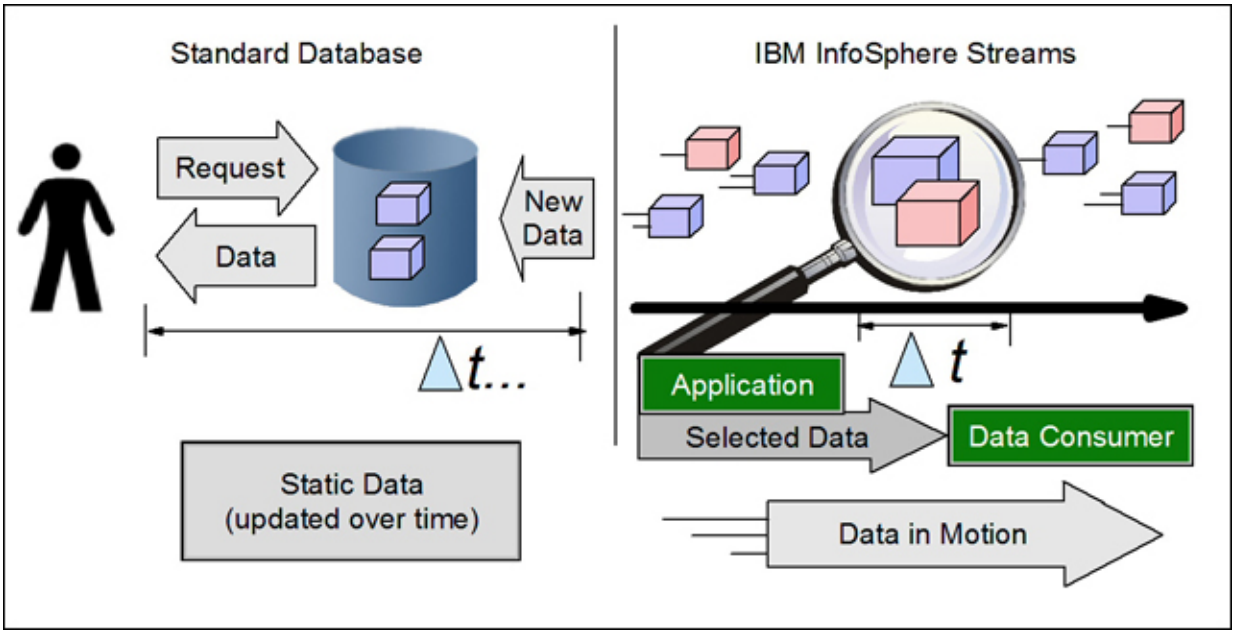


Figure 3. A standard relational database compared to IBM InfoSphere Streams

In the InfoSphere Streams approach (right side in Figure 3), the data of interest within the stream is defined and then processed as it flows by the application. The data is processed while it is in motion and is typically not stored.

Because InfoSphere Streams applications are always running and continually sending selected data that meets the defined criteria to users, they are well suited for answering *always-on* or *continuous* types of questions. For example, such questions might ask what the rolling average is, how many parts have been built so far today, or how production at a specific time today compares with that same time yesterday. The questions might include other continuous, subsecond response time statistics. Because of the nature of the questions that InfoSphere Streams can answer, it provides join-like capabilities that are beyond the capabilities in standard SQL.

The InfoSphere Streams runtime system as shown in Figure 4 is a collection of components and services.

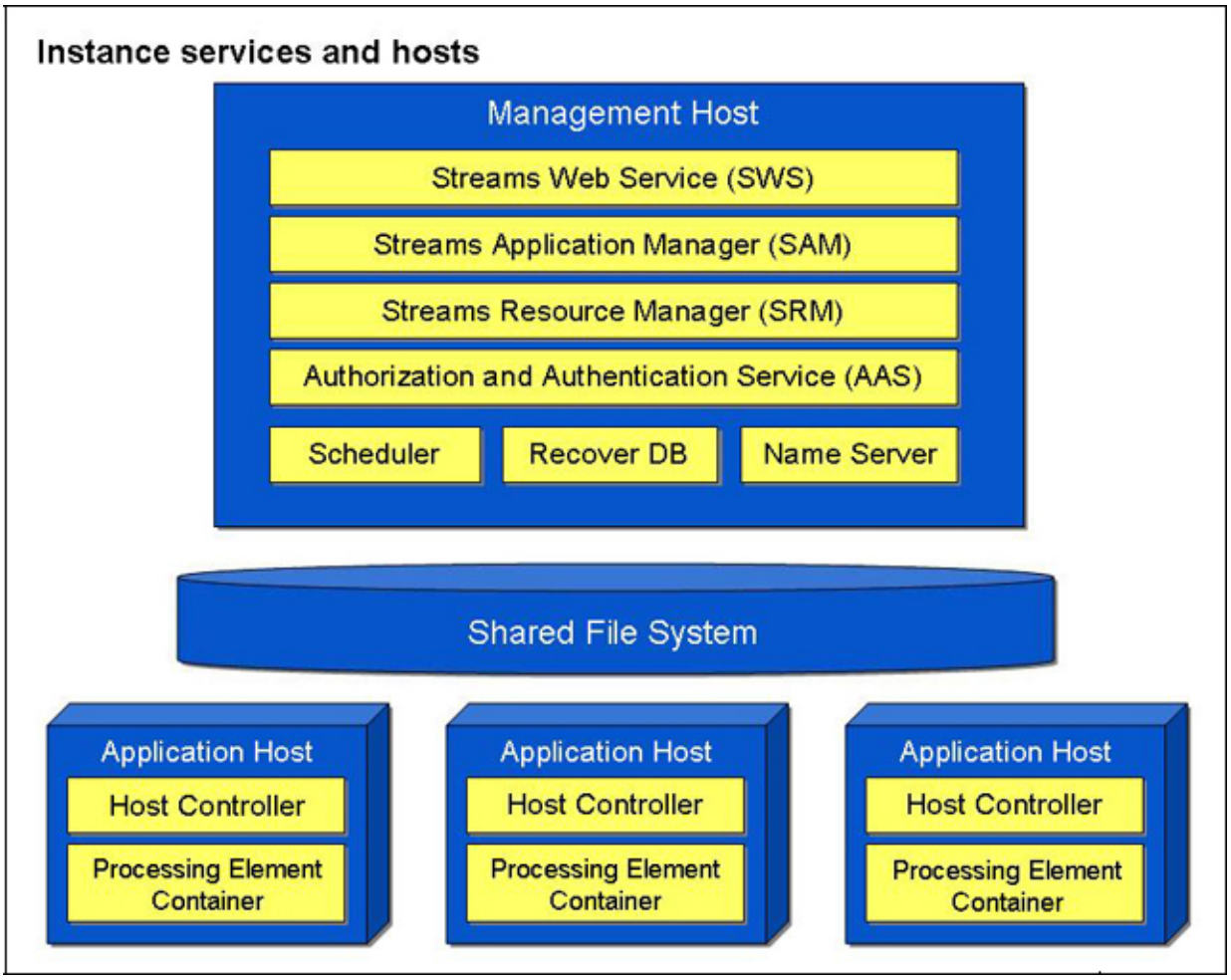

Figure 4. InfoSphere Streams runtime components

These components and services work together to provide scalability and capability on high volumes of streaming data. The runtime system continuously monitors the state and utilization of its computing resources. When applications are running in the InfoSphere Streams environment, they can be dynamically monitored across a distributed collection of hosts by using the InfoSphere Streams Studio.

## Usage scenarios

Throughout its research and development phase, InfoSphere Streams demonstrated success by delivering cutting edge commercial and scientific applications. Many of these applications have shown clients a way to capture the formerly unattainable competitive edge as illustrated in the following examples.

### Telecommunications: Mediation and analytics of the call detail record

The challenge of closing certain technology and business gaps is especially apparent for cellular service providers in Asia. Chips that are embedded in cell phones enable email, texting, pictures, videos, and information sharing by using social sites such as Facebook. For each phone call, email, web browse, or text message, cellular phone switches emit call detail records (CDRs). To ensure that no data is lost, the switches emit 2 CDRs for each transaction, which must later be deduplicated for the billing support

systems. A rising volume of data caused mediation of CDRs to become more difficult to perform in a timely manner. Phone number portability enabled subscribers to move to a competitor at any time. Some sophisticated users even switch between multiple providers at different times during a day to take advantage of certain promotions. Providers needed to reduce the window for processing CDRs to near real time and to perform real-time analytics in parallel, to predict which customers might leave for a competitor, known in the industry as *churn*. With this real-time insight into customer behavior, providers can take action to retain a higher percentage of customers.

By using InfoSphere Streams, with its agile programming model, customers can handle their huge volume of CDRs with low latency, while providing churn analysis on the data at the same time. At one company, a peak rate of 208,000 CDRs per second is processed with an end-to-end processing latency of under 1 second. Each CDR is checked against billions of existing CDRs, with duplicates eliminated in real time, effectively cutting in half the amount of data that is stored in their databases.
This example illustrates a key InfoSphere Streams use case of simultaneous processing, filtering, and analysis in real time. High availability, automated fault tolerance and recovery, in addition to real-time dashboard summaries are also currently in use and improving IT operations. Real-time analysis of CDRs is leading to improved business operations by performing such tasks as churn prediction and campaign management to improve the retention rate of their current customers.

Figure 5 illustrates the flow of the CDR processing architecture with InfoSphere Streams.

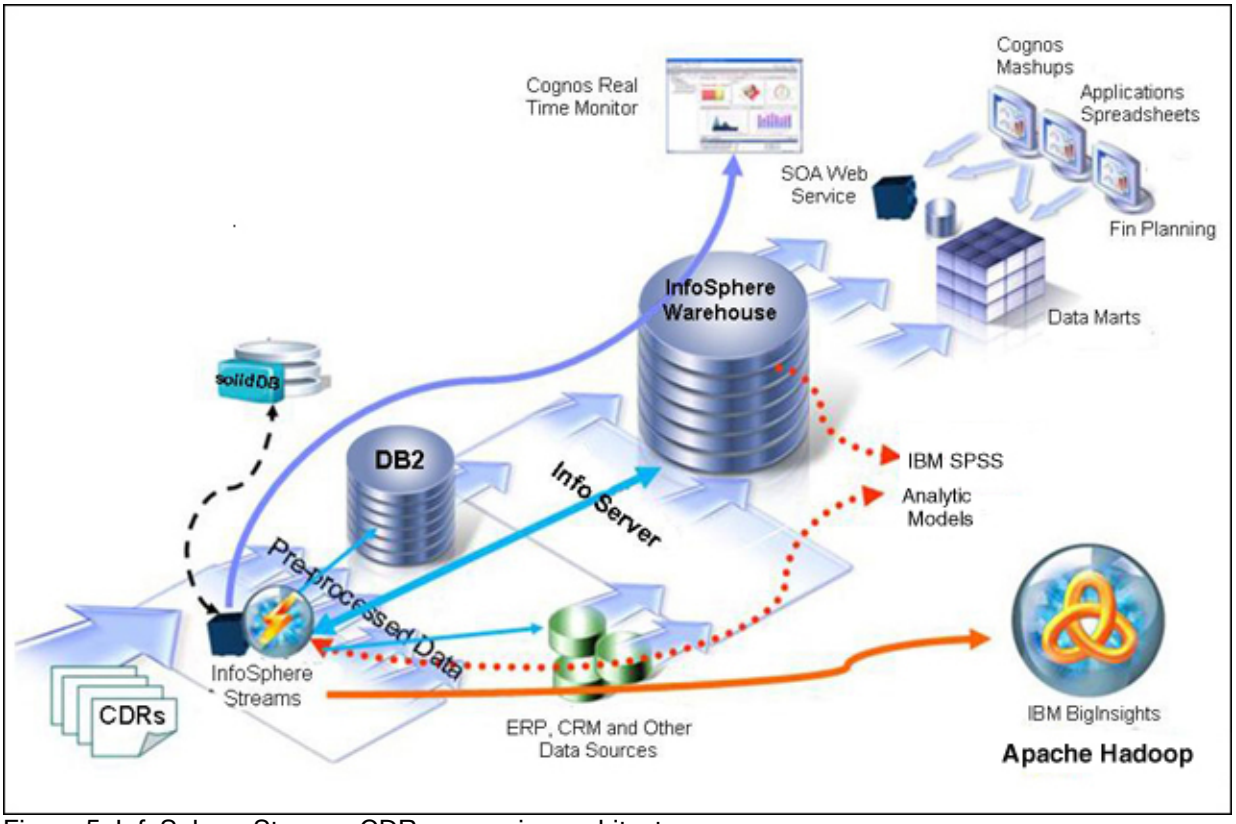

Figure 5. InfoSphere Streams CDR processing architecture

### Financial services: Finding arbitrage opportunities faster than the competition

Many segments of the financial services industry rely on rapidly analyzing large volumes of data to make near real-time business and trading decisions. Today these organizations routinely consume market data at rates exceeding one million messages per second, twice the peak rates they experienced only a year ago. This dramatic growth in market data is expected to continue for the foreseeable future, outpacing the capabilities of many current technologies. Industry leaders are extending and refining their strategies by

including other types of data in their automated analysis. Sources range from advanced weather prediction models to broadcast news. The development of IBM InfoSphere Streams is based on a trading prototype running on a single 16 core x86 computer that could processing OPRA data feeds at up to 5.7 million option messages per second, with latencies of under 30 microseconds. After recompiling the application, InfoSphere Streams can scale even further when deployed across a cluster of computers.

#### Health monitoring: Predicting the onset of illness earlier

Stream computing can be used to better perform medical analysis and reduce workload on nurses and doctors. Privacy-protected streams of medical device data can be analyzed to detect early signs of disease, correlations among multiple patients, and efficacy of treatments. A strong emphasis is on data provenance within this domain. *Provenance* is the tracking of how data are derived as it flows through the system. As an example of stream computing in the healthcare industry, a hospital in Toronto, Canada uses InfoSphere Streams technology to monitor premature babies in a neonatal intensive care unit to help predict the onset of illness. Remote telemetry from a US hospital has been operational for a year using the same analytic routines. And earlier this year, more hospitals in China and Australia began implementing this solution. For more information, see *University of Ontario Institute of Technology: Leveraging key data to provide proactive patient care* at:
http://public.dhe.ibm.com/common/ssi/ecm/en/odc03157usen/ODC03157USEN.PDF

## Integration

The IBM solution to handle big data includes several platforms as illustrated in Figure 6.

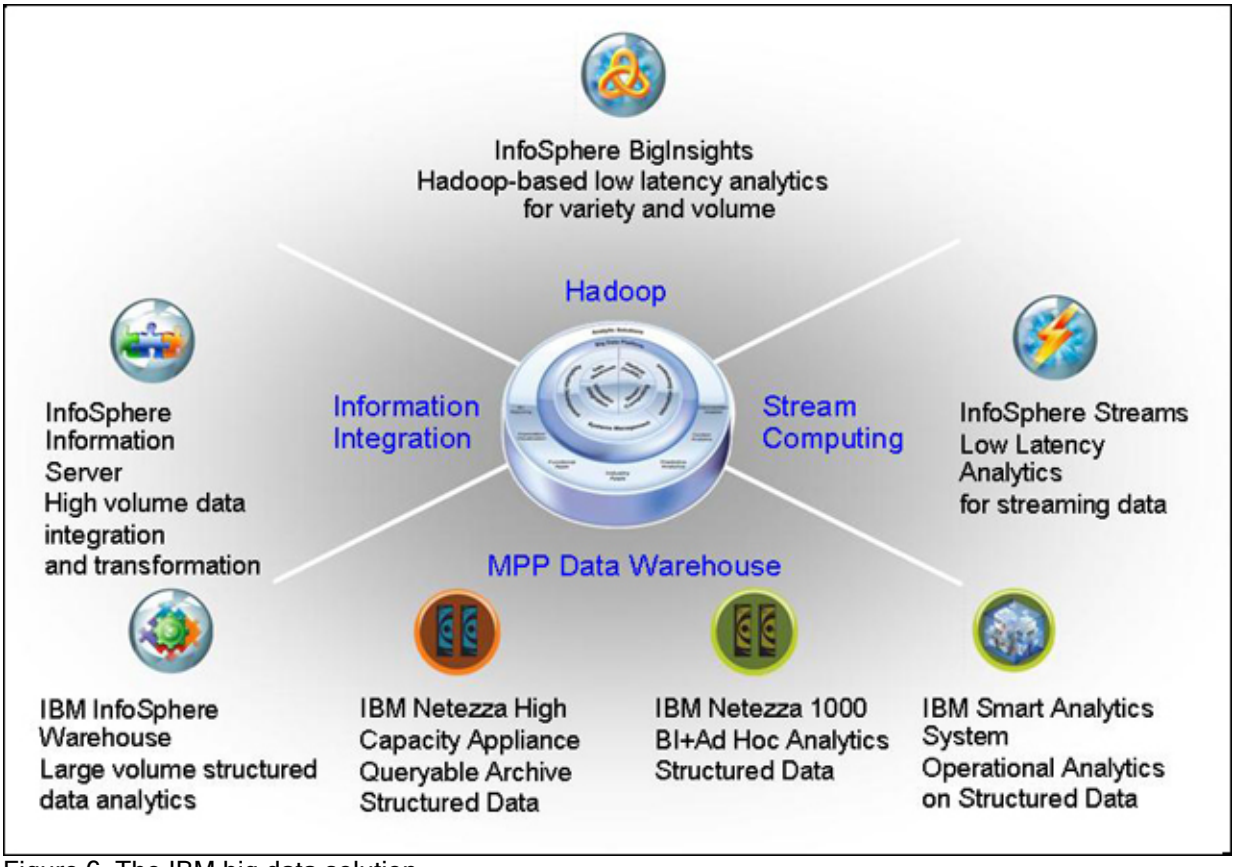

Figure 6. The IBM big data solution

The IBM big data solution includes the following platforms:

- IBM InfoSphere BigInsights™

  After storing the raw data in InfoSphere BigInsights, firms can manipulate, analyze, and summarize the data to gain new insights and to feed downstream systems.

- IBM InfoSphere Information Server

  By using InfoSphere Information Server for data integration, you can transform data in any style and deliver it to any system, ensuring faster time to value.

- IBM InfoSphere Data Warehouse

  InfoSphere Warehouse represents the IBM offering for implementing integrated business intelligence (BI) solutions. The BI framework enables the transformation of the data warehouse from a static repository, primarily used for batch reporting, into an active end-to-end platform for BI solutions.

- IBM Netezza®

  The Netezza appliances integrate database, processing, and storage in a compact system that is optimized for analytical processing and that is designed for flexible growth.

- IBM Smart Analytics System

  Smart Analytics System, which takes advantage of the appliance architecture, is a preintegrated system that features a powerful data warehouse foundation and extensive analytic capabilities.

## Supported platforms

- X86 32-bit or 64-bit systems with a minimum of 500 MB of memory
- 2 GB of memory to run simple applications, such as the Commodity Purchasing Sample Application that is included with InfoSphere Streams

The IBM System x3550 M4 reference architecture is supported.

## Ordering information

This product is available only through IBM Passport Advantage®. It is not available as a shrink wrapped product.

- License function title: InfoSphere Streams
- Product group: IBM InfoSphere

Table 1 shows the ordering information.

Table 1. Ordering part numbers and feature codes

| Program name | PID number | Charge unit description |
|---|---|---|
| IBM InfoSphere Streams | 5724-Y95 | Per Resource Value Unit |

## Related information

For more information, see the following documents:

- *IBM InfoSphere Streams V3.0: Addressing volume, velocity, and variety*, SG24-8108
  http://www.redbooks.ibm.com/abstracts/sg248108.html
- Big data
  http://www.ibm.com/software/data/bigdata
- IBM InfoSphere Streams V3
  http://www.ibm.com/software/data/infosphere/streams
- IBM Offering Information page (to search on announcement letters, sales manuals, or both):
  http://www.ibm.com/common/ssi/index.wss?request_locale=en

  On this page, enter `InfoSphere Streams` V3, select the information type, and then click **Search**. On the next page, narrow your search results by geography and language.
- IBM InfoSphere Streams Information Center
  http://publib.boulder.ibm.com/infocenter/streams/v2r0/index.jsp

# Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:** INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document was created or updated on January 30, 2013.

Send us your comments in one of the following ways:

- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbook@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips0948.html .

# Trademarks